BEFORE THE
FEDERAL COMMUNICATIONS COMMISSION
Washington, D.C. 20554

In the Matter of	)	
	)	
Application of BellSouth Corporation,	)	
Pursuant to Section 271 of the	)	CC Docket No. 01-277
Telecommunications Act of 1996	)	
To Provide In-Region, InterLATA Services	)	
In Georgia and Louisiana	)	

DECLARATION OF ROBERT M. BELL
ON BEHALF OF AT&T CORP.

October 19, 2001

Table of Contents

Page

BEFORE THE
FEDERAL COMMUNICATIONS COMMISSION
Washington, D.C. 20554

In the Matter of	)	
	)	
Application of BellSouth Corporation,	)	
Pursuant to Section 271 of the	)	CC Docket No. 01-277
Telecommunications Act of 1996	)	
To Provide In-Region, InterLATA Services	)	
In Georgia and Louisiana	)	

DECLARATION OF ROBERT M. BELL
ON BEHALF OF AT&T CORP.

I. INTRODUCTION AND QUALIFICATIONS.

1. My name is Robert M. Bell. My business address is AT&T Labs-Research, 180 Park Avenue, Florham Park, New Jersey 07932.

2. I received a Ph.D. in Statistics from Stanford University in 1980. From 1980 to 1998, I was promoted to Senior Statistician at RAND, a non-profit institution that conducts public policy analysis. While at RAND, I supervised the design and/or analysis of many projects, including large multi-site evaluations in the fields of preventive dentistry, drug prevention, and depression care. I also headed the RAND Statistics Group from 1993 to 1995 and taught statistics in the RAND Graduate School from 1992 to 1998. In 1998, I joined the Statistics Research Department at AT&T Labs-Research, where I am a Principal Member of Technical Staff.

3. I have authored or co-authored fifty articles on statistical analysis that have appeared in a variety of refereed, professional journals. I am a fellow of the American Statistical Association. I am currently a member of the Committee on National Statistics organized by the National Academy of Sciences as well as the Academy's

Panel to Review the 2000 Census. I have attached a copy of my curriculum vitae as Exhibit RMB-1.

II. PURPOSE OF DECLARATION

4. The purpose of this declaration is to discuss statistical issues related to the Louisiana and Georgia Performance Measurement plans and the *BellSouth Telecommunications, Inc. OSS Evaluation – Georgia, Master Test Plan Final Report* ("Final Report") on the third-party test conducted in Georgia. I describe several problems with the data, analysis, and conclusions reported in the third-party test Final Report. In particular, I explain that there is insufficient evidence to support many of the conclusions reached by KPMG Consulting, Inc. ("KCI").

5. I also comment on issues relating to BellSouth's performance measures plan generally. I explain how the truncated z statistic can conceal discrimination when used as it is in the Georgia and Louisiana plans. I explain why AT&T believes that the value of delta used to compute balancing critical values in the Performance Measurement plan should be 0.25 or lower for both Tier I and Tier II tests. I also describe why the calculation of "affected volume," which is central to the remedy calculations, is inappropriate.

III. THIRD-PARTY TEST

6. Based on my review of the Final Report, I have concluded the following:

- KCI should not have applied a statistical analysis known as "P-value" analysis to tests involving Commission-established benchmarks. This process rewarded BellSouth with a "satisfied" rating even though BellSouth failed to meet the benchmark.
- In cases where it may have been appropriate to apply some sort of statistical analysis to account for random variation in results, KCI applied

a P-value analysis to account for random "bad" results (to reach a "satisfied" result) but did not use a statistical analysis to rule out random "good" results.

- KCI's "satisfied" determinations based on aggregated rather than disaggregated service types masked poor performance that otherwise would have led to a conclusion of "not satisfied."

- Preferential treatment of CLEC orders from Georgia during the test period invalidate conclusions about any measures involving partially mechanized or non-flow-through orders.

- The absence of blindness to the test subject allowed BellSouth to give systematic preferential treatment to the tested orders. Consequently, results from approximately 30% of the orders submitted may be biased and are therefore invalid.

- KCI should have implemented military style testing in a manner that revealed BellSouth's true performance. When KCI retested following a "not satisfied" result, KCI often applied a less robust, less reliable retest using a smaller sample size.

- KCI's practice of using its "professional judgment" to change a finding of "not satisfied" to "satisfied" is unusual and questionable.

In short, the conclusions drawn by KCI are based on incomplete statistical analysis. More complete analysis shows that many of the conclusions that particular standards were satisfied are not justified by the data in the test.

A. Statistical Analysis of Test Results.

7. Some background information regarding the standards used for the test is necessary to understand my concerns regarding the statistical analysis of test results. KCI used three types of standards: (1) parity with a retail analog; (2) benchmarks, *i.e.*, quantitative standards set by either the Georgia Public Service Commission, BellSouth, or KCI; and (3) subjective standards set by KCI. Because KCI applied statistical methods only to the objective standards (1) and (2) above, I will comment only on those.

8. When reviewing the analysis of test results against these two standards, it is useful to consider three quantities that share a common scale: (1) the observed service, (2) the true service, and (3) the standard. The definition of each of these quantities differs slightly depending upon whether the standard is a benchmark or parity.

9. For benchmarks, the truth, an unknown quantity that characterizes the service process, is the mean or average that would occur if we could observe an unlimited number of pseudo-CLEC observations under the same conditions as the test. The observed service is simply the mean (average) measurement or the proportion of successes observed for all pseudo-CLEC cases in the test. The observed service estimates the true service based on a finite sample of observations. The benchmark standard is a specified quantity, *e.g.*, a 95% success rate.

10. Parity measures involve a comparison of the results or process available for BellSouth retail with the results or process provided for the pseudo-CLEC. For parity measures, the truth (or true service) is the difference between the values of the mean (or proportion) for unlimited pseudo-CLEC service and the mean (or proportion) for unlimited retail results. Similarly, the observed service for parity purposes is the difference between the mean (average) measurement or the proportion of successes observed for the pseudo-CLEC cases and the corresponding value for BellSouth's retail cases. For parity measures, the standard is always a difference of zero.

1. KCI Should Not Have Applied Statistical Analysis To Results Measured Against Commission-Established Benchmarks

11. KCI's statistical analyses (P-values) for benchmarks assume that the standard (*e.g.*, 95% success) was intended as a standard for the true service.

However, BellSouth and AT&T agreed in Louisiana, Georgia, and other states that benchmarks should be treated as strict cutoffs for all samples of 30 or larger. In response to Louisiana's Stipulated Open Issue 19 ("What is the appropriate adjustment for small samples sizes when a benchmark is used to determine the standard of performance?"), BellSouth stated, "Benchmarks for Proportion measures are an exact statement of the commitment (*i.e.*, "X%")." It continued, "For both types of benchmarks, adjustments will be made for small sample sizes (with the exception of Collocation for which no adjustments will be made). Small sample size adjustments was a concept proposed by AT&T and subsequently adopted by BellSouth for samples ranging from 5 to 30." (*BellSouth's Comments to Stipulated Open Issues*, before the Louisiana Public Services Commission, Docket No. U-222523, Subdocket C, March 20, 2000). Based on this interpretation, it was improper for KCI to use statistical analysis to reclassify benchmark standards as satisfied. Adjustments for sample sizes of 30 or less should not have been an issue in the Georgia test because KCI controlled the sample sizes. (*See infra.* ¶¶ 32-35.)

12. The Georgia Public Service Commission (GPSC) set the benchmark standards for the following twenty-nine individual tests in the third-party test: PRE-1-1-1, PRE-4-1-1, PRE-5-1-1, O&P-1-1-1, O&P-1-3-2a, O&P-1-3-2b, O&P-1-3-3a, O&P-1-3-3b, O&P-1-3-5, O&P-2-1-1, O&P-2-3-2a, O&P-2-3-2b, O&P-2-3-3a, O&P-2-3-3b, O&P-3-1-1, O&P-3-1-2, O&P-3-3-3, O&P-3-3-4, O&P-4-1-1, O&P-4-1-2, O&P-4-2-1, O&P-4-2-2, O&P-4-3-3, O&P-4-3-4, O&P-5-2-3, O&P-10-1-1, O&P-10-1-2, O&P-10-3-3, and O&P-10-3-4. Consequently, statistical analysis (P-values) should not be

applied to any of these benchmarks. Any benchmark for which the observed service does not meet or exceed the standard should be classified as not satisfied.

2. KCI Did Not Conduct A Complete Statistical Analysis Of Results

13. Based on the way KCI structured the test, "satisfied" in the Final Report does not mean that BellSouth met the specified standard. Instead, "satisfied" implies only that there was not enough evidence to conclude that the true service process was below standard. In light of the small sample sizes in the test, BellSouth could easily have received a score of satisfied despite seriously substandard performance.

14. KCI's analysis of test results began with the null hypothesis.[1] The null hypothesis assumes that BellSouth is performing at the standard.

15. If the observed results in a specific test met or exceeded the standard, KCI did no further analysis and classified the standard as "satisfied." Whenever the observed results for a measure failed to reach the specified standard, KCI applied one of four statistical tests to compute a P-value. The P-value compares the observed results with the standard and provides a quantitative measure of how likely a result as bad as the one observed would be under the null hypothesis that the true service process employed by BellSouth exactly met the standard. In other words, the P-value indicates how reasonable it is to conclude that the observed result is explained solely by bad luck as opposed to any deficiency in the true service. Small P-values provide

[1] The null hypothesis takes one of two forms. For parity measures, the null hypothesis states that the distribution of outcomes is the same for CLEC and BellSouth customers. For benchmarks, KCI's null hypothesis is that the mean or proportion for CLEC customers is exactly at the standard. In my view, it is inappropriate to apply statistical analysis to benchmarks at all. If it is applied, however, the guidelines discussed in this section should be followed.

evidence against the null hypothesis that the true process was meeting the standard. They represent a smaller chance that the observed result was due to chance.

16. KCI used the following rule: If BellSouth's observed performance did not meet the specified standard, KCI calculated a P-value. If the P-value was less than 0.05, KCI concluded that the standard was "not satisfied." If the P-value exceeded 0.05, KCI concluded that the standard was "satisfied."

17. For example, Test O&P-5-2-3 measured whether provisioning was completed on time for Coordinated Customer Conversion orders. The BellSouth Service Quality Measurements Plan applies a standard of 95% within 15 minutes of the scheduled start time for coordinated customer conversions. This means that, to meet the standard, BellSouth would have to start coordinated customer conversions within 15 minutes of the scheduled start time for 95% of the coordinated conversions it performs. Out of 63 observed conversions with CLECs, there were 57 successes, a success rate of 90.4%. Because the observed rate failed to meet the standard, KCI calculated a P-value. Based on a P-value of 0.0945, KCI concluded that the 95% standard was "satisfied." In essence, KCI based its conclusion of "satisfied" on a 0.09 probability that this bad a result could occur by chance if BellSouth is meeting the standard.

18. The possibility that substandard results would be observed when the null hypothesis is true, *i.e.*, BellSouth's service meets the standard, is called a Type I error. KCI's rule, that if the P-value exceeded 0.05 the standard was satisfied, explicitly controls the probability of Type I error to be 0.05 or less. In other words, under KCI's rule, for any measure where BellSouth's true service process meets the standard, there is

at most a 1-in-20 chance that BellSouth's observed service would be found out of compliance.

19. KCI's procedure asks the question: "Given the observed results, how good could the true service be?" However, KCI does not address the corresponding question: "Given the observed results, how bad could the true service be?" A balanced, complete analysis needs to address both questions.

20. In order to answer the latter question – Given the observed results, how bad could the true service be? – Type II errors must be considered. A Type II error occurs when an *alternative hypothesis* is true (*i.e.*, BellSouth's true process fails to meet the standard by a certain amount), but the conclusion is that the standard was satisfied. Type II errors are important because they are instances in which BellSouth's process is not up to the standard but KCI concludes that it is. Consequently, it is just as important to limit the probability of Type II errors as it is to limit Type I errors. In response to an interrogatory about how it "set sample sizes to protect against Type II errors," KCI provided no evidence that it had considered Type II errors when setting sample sizes.

21. With KCI's fixed cutoff of 0.05 for P-values, sample size determines the probability of a specific Type II error. With sufficient sample size for a measure, Type II errors do not pose much problem. Indeed, KCI's witness agreed that "as the sample size increases you reduce the probability of observing type two errors." *See* Transcript of Deposition of Alan J. Salzberg, September 24, 2001 ("Salzberg Tr.") 96:13-16 (RMB-2).

22. For example, consider measure O&P-5-2-3 again, for which the standard is 95%. If we used a sample size of 250 and BellSouth's true performance is

equal to 90.0%, the probability of a Type II error would be just 0.12. That is, with an adequate sample size, the odds would strongly favor seeing the failure in the observed service.

23. Type II error presents a serious problem, however, for the sample sizes generally employed in this test. With the sample of 63 used for O&P-5-2-3, the probability of a Type II error when the true success rate is 90.0% equals 0.71. In other words, if BellSouth was failing at twice the rate specified in the standard (*i.e.*, 90.0% success, 10.0% failure), the odds are more than two-to-one that KCI would have judged the standard to be satisfied.

24. In fact, there would be a 0.25 probability of a Type II error even if BellSouth's true performance rate had dropped to 85.0%. In other words, even if BellSouth were performing at 85.0%, one out of four times KCI would incorrectly classify the standard as "satisfied." The test should have been designed to avoid such large Type II error probabilities for serious violations of this sort.

25. Based on these large Type II error probabilities, it is clear that further analysis is required before concluding that any standard in the test is satisfied.

26. The best way to address both Type I and Type II errors is to compute a two-sided confidence interval that summarizes the uncertainty associated with observed performance. For example, the exact, two-sided 90-percent confidence interval for measure O&P-5-2-3 runs from an upper limit of 95.8% down to a lower limit of 82.1%. This confidence interval states the range of true proportions that are consistent with the observed results: 57 successes out of 63.

27. In other words, having observed these results, BellSouth's true performance is likely to be anything from 82.1% to 95.8%. For proportions outside the confidence interval (above 95.8% or below 82.1%), the observed number of successes is either surprisingly low or surprisingly high. By surprising, I mean that we would expect such an extreme result in either direction less than 5 percent of the time. Because the confidence interval is two-sided, the coverage probability is 90 percent (100 percent minus 5 percent off each end).

28. Two-sided confidence intervals provide information about both how good and how bad the true level of service might be. The upper confidence limit tells the same story as KCI's P-value. The fact that the upper confidence limit for O&P-5-2-3 exceeds 95% is equivalent to the P-value exceeding 0.05.

29. However, the lower confidence limit provides important information that the P-value misses. Indeed, the lower confidence limit of 82.1% for O&P-5-2-3 tells a very different story from the P-value alone. The lower confidence limit means that the observed data cannot rule out the possibility that the true failure rate was as high as 17.9%, *i.e.*, that it exceeded the standard by a factor of 2.0, 3.0, or even 3.6. Consequently, KCI's conclusion of "satisfied" could easily have been a serious Type II error. Indeed, KCI's witness acknowledged that "if BellSouth's observed performance is at 90.4 percent . . .their true performance could be worse than that." Salzberg Tr. 78:14-18 (RMB-2). Clearly, it is wrong to conclude that BellSouth's service process meets the standard for this measure.

30. Rather than classify O&P-5-2-3 as satisfied, KCI should have classified it as "inconclusive" because neither standard nor significantly substandard performance could be ruled out.

31. This problem with KCI's analysis can be corrected. Two-sided 90-percent confidence intervals should be computed for all measures. In addition, a threshold should be specified indicating a level of poor service that must be ruled out before the standard is classified as satisfied. BellSouth's performance on a measure would be classified as satisfied only if the upper confidence limit exceeded the standard *and* the lower confidence limit exceeded the threshold. If the confidence interval is large enough to cover both values, the result would be classified as *inconclusive.* For a measure with a standard of 95 percent, I believe that 87.5% would be an appropriate threshold. This threshold represents a failure rate that is 2.5 times that specified in the standard. Observed data that are consistent with an even greater lack of compliance should not be used as evidence that the standard was satisfied. If my proposed procedure and threshold is applied to O&P-5-2-3, this test result would be classified "inconclusive."

32. If the sample is small enough, the lower confidence limit may fail to meet the threshold even when the observed result exceeds the standard. Consequently, confidence intervals should be computed and compared with a threshold even when KCI did not compute P-values.

33. For any measure with inconclusive results, more data should be collected. Sufficient additional data should be collected and combined with the existing data to narrow the confidence limits substantially before recomputing a P-value and

confidence limits. If the new confidence limits meet the conditions outlined above, the measure can be determined to be satisfied or not satisfied.

34. The problem illustrated by the inadequate sample size of O&P-5-2-3 is not an isolated instance. Similar sample sizes occurred for the following thirteen measures: PRE-1-3-1 (n=57), PRE-1-3-2 (n=68), PRE-1-3-3 (n=73), PRE-1-3-5 (n=51), PRE-1-3-9 (n=83), O&P-1-3-2b (n=70), O&P-1-3-3a (n=50), O&P-1-3-3b (n=50), O&P-2-3-2a (n=89), O&P-2-3-3b (n=74), O&P-5-2-1 (n=89), O&P-5-2-2 (n=72), and O&P-5-2-5 (n=55).

35. Small sample size is especially a concern for five of these measures that were classified as satisfied on the basis of the statistical tests: PRE-1-3-1, PRE-1-3-2, O&P-2-3-2a, O&P-5-2-2, and O&P-5-2-5. For example, the confidence intervals for O&P-5-2-2 (n = 72) and O&P-5-2-5 (n = 55), both 95% benchmarks, look similar to that for O&P-5-2-3 and would also lead to classifications of "inconclusive."[2] Measure O&P-2-3-2a, timeliness of response for fully mechanized order error notices, had a GPSC-set benchmark of 97% within one hour. The observed success rate was 94% (84 of 89). KCI classifies this standard as satisfied even though the 90-percent confidence interval ranges from 97.8% all the way down to 88.6%. This lower limit would imply almost four times the failure rate specified by the GPSC.

36. Clearly, each of these measures should lead to a classification of "inconclusive," not a classification of "satisfied." The other measures listed above, both those that were satisfied with use of the P-values and some that were satisfied without a P-value, are also suspect because of the small sample sizes tested.

[2] The reference to "n" refers to the sample size for the test.

37. An even greater problem exists with some of the very small sample sizes, for example O&P-1-3-5 (n=9), O&P-1-3-6 (n=15), and O&P-2-3-6 (n=15). For these sample sizes, even 100% observed performance at the standard would be inconclusive. For example, on measure O&P-2-3-6, a 95% benchmark, BellSouth achieved 15 successes in 15 tries. Nevertheless, the lower confidence limit is only 81.9%. As KCI's witness acknowledged, "[t]he smaller the sample size, the less precise, in general the less precise your results are going to be." Salzberg Tr. 87:17-19 (RMB-2). A sample size of 15 is simply inadequate to rule out the real possibility that true service is far below the standard.

38. Despite the generally inadequate sample sizes, my proposed procedure would substantiate KCI's conclusions whenever the observed service is good enough to rule out true service as bad as the threshold.

3. Disaggregation.

39. Evaluation of BellSouth's provisioning of individual service and activity types requires analysis of data for each individual service and activity. KCI did not adequately test disaggregated service types. Although KCI's Final Report provides test results in tables broken down by the listed service and activity types, conclusions were reached based on aggregated results. Examination of the tables reporting the disaggregated results reveals two problems with KCI basing its decisions on aggregated results. First, in certain instances, KCI reached a conclusion of satisfied when BellSouth's performance at the service/activity level was well below the standard for certain service types. Second, due to the small sample sizes for the individual service/activity types, most results would be inconclusive.

40. KCI determined that BellSouth satisfied test O&P-1-3-2b (BellSouth's EDI interface provides timely partially mechanized order clarifications), even though BellSouth did not meet the specified standard for 2-wire loops with local number portability. (*See* Final Report, p. V-A-34). In the first retest, KCI evaluated 34 orders for 2-wire loops with LNP. On those transactions, BellSouth failed to meet the Commission's standards for order clarification and error notices for either fully mechanized or partially mechanized orders. BellSouth completed just 8 of 14 partially mechanized orders in less than 24 hours—significantly less than the GPSC-approved standard of 85% (P-value = 0.012). *Id*. Nonetheless, based on the summary data for partially mechanized orders across all service types, KCI determined that BellSouth has satisfied the Commission's standard for timely error and clarification notices for partially mechanized orders. Accordingly, KCI concluded that BellSouth had satisfied the test even though the test results reveal that BellSouth did not satisfy the Commission's standard for timely order and clarification notices for orders that allow a customer to keep his or her own phone number when switching carriers.[3]

41. The Louisiana Public Service Commission ordered the same standard for Performance Measurements as the Georgia Commission. (*See* General Order in Docket U-22252, sub-docket C, Exhibit B to Attachment A (May 14, 2001)). Accordingly, BellSouth did not meet the standards of the Louisiana Commission either.

42. In other situations, KCI determined that BellSouth satisfied the standard without testing any of certain important service/activity types. Test 1-3-3a and 1-3-3b tested firm order confirmation ("FOC") timeliness. As with the previous example,

[3] KCI only tested one standalone LNP order on this retest.

KCI did not perform its evaluation at the required levels of disaggregation, nor did it set its sample sizes to ensure adequate evaluation of each service/activity type. When evaluating FOC timeliness, KCI initially tested fifteen 2-wire loops with LNP and three LNP standalone orders. (*See id.* at V-A-41-42.) In the first retest, KCI evaluated twenty-six 2-wire loops with LNP and fourteen standalone LNP orders. (*See id.* at V-A-43-46.) After this first retest, based on summary data aggregated across all service/activity types, KCI determined that BellSouth had met the Commission standard of 85% of FOCs returned within thirty-six hours for orders that did not flow through. The disaggregated view, however, reveals that no orders for 2-wire loops with LNP and no orders for standalone LNP were included in this evaluation of non-flow-through orders. Thus, again, this test does not reveal that BellSouth has satisfied the Commission's standards for FOC timeliness for non-flow-through orders when the orders are ones that permit customers to keep their own phone numbers when switching carriers.

43. While the two samples above present the most extreme examples, virtually none of BellSouth's sample sizes were adequate to test BellSouth's performance at the disaggregated level. For example, on tests O&P-1-3-2a and O&P-1-3-2b, which tested whether BellSouth provided timely order error notices for both fully mechanized (O&P-1-3-2a) and partially-mechanized (O&P-1-3-2b) orders, BellSouth tested small numbers of each disaggregated service type. In the first retest for fully mechanized orders, KCI tested only twenty-five 2-wire loops-design, only twenty 2-wire loops-nondesign, only thirteen 2-wire loops with LNP-design, only seven 2-wire loops with LNP-nondesign, only three switch ports, and only eleven loop port combinations. As explained in Section I.A.2, sample sizes like these are not sufficient to support the

conclusion that the standard has been met even when there are few or no observed failures.

44. For partially mechanized orders, BellSouth tested only twenty-three 2-wire loops-design, only six 2-wire loops-nondesign, nine 2-wire loops with LNP-design, five 2-wire loops with LNP-nondesign, one LNP stand-alone, five switch ports, and seven loop port combinations. These sample sizes are similarly too small to have any statistical power. Anyone analyzing the observed results would not have nearly enough evidence to predict BellSouth's true performance.

45. KCI also based its determinations upon aggregated test results for other tests, including O&P-1-3-3a, O&P-1-3-3b, O&P-2-3-2a, O&P-2-3-2b, and O&P-2-3-3b. Each of the problems addressed above applies to each of these tests. If BellSouth is failing to meet the standard for some services or activities, an assessment based on aggregated data could easily miss the failure. Moreover, certain activity types were not tested at all. The sample sizes for other activity types are too small to support any valid conclusions about the individual service types. Consequently, the Commission lacks the information it needs to determine whether BellSouth is meeting the standard for individual services and activities.

B. Issues Affecting Data Validity.

1. Representativeness.

46. The validity of the conclusions of a statistical analysis can be no better than the data that go into the analysis. BellSouth has asked the Commission to use the results of the Georgia test as evidence about the OSS support available to CLECs in both Georgia and Louisiana. A key question underlying the validity of the third-party test for this purpose is whether the service received by the pseudo-CLEC established by

KCI is representative of service that would have been received by real CLECs in those two states. If, for any reason, this service were substantially better, the conclusions reached by KCI would be completely invalid. KCI's witness agreed that a "random representative sample makes a test more persuasive than a sample that is not necessarily representative." Salzberg Tr. 95:12-15 (RMB-2).

47. Documents recently produced by PricewaterhouseCoopers reveal that during the Georgia and Florida tests, BellSouth's two Local Carrier Service Centers ("LCSCs") had a policy of giving preferential treatment to Local Service Requests ("LSRs") received for manual processing from Georgia and Florida. This practice apparently was not disclosed to KCI or the commissions, and it continued through November 2000 in the Atlanta LCSC and through April 2001 in the Birmingham LCSC. As a result, for the majority of the Georgia test, LSRs from KCI received special handling that could be expected to improve BellSouth's perceived performance.

48. BellSouth's policy of preferential treatment introduced serious bias that taints results from all partially-mechanized and non-mechanized orders processed as part of the test. On average thirty percent of orders evaluated during the period of the third-party test required manual processing. For all measures involving such orders, the test results are completely invalid.

49. Even if orders from Georgia had received no special handling during the test, KCI's conclusions would be invalid if pseudo-CLEC orders were treated differently from other CLEC orders during the period of the test. KCI devotes just two short paragraphs of the Final Report to this important concern under the heading "*Blindness*" (Section II-6.5).

50. Blindness means that the subject of an experiment does not know whether he or she is in the treatment or control condition. The most common example of this occurs in clinical trials. Patients who believe that they are receiving an experimental treatment may tend to improve simply because of that belief. This is why control-group patients routinely receive placebos.

51. In this test, blindness refers to BellSouth. It means that BellSouth should have been unable to distinguish service requests of the pseudo CLEC from requests of any other CLEC. That is, BellSouth should not have known which service requests it was being evaluated on. In the absence of blindness, it is impossible to establish that the results observed by KCI are representative of the service that real CLECs were receiving at the same point in time.

52. Unfortunately, the test was not designed in a way to blind BellSouth to the identity of KCI orders. The report acknowledged, "Yet, it was virtually impossible for the KCI/HP test to be truly blind to BellSouth." While complete blindness was probably impossible to achieve, every effort should have been made to minimize the opportunity for BellSouth to discover the source of the service requests. Instead, as KCI reports, "Each CLEC has a unique set of IDs assigned by BellSouth that must be included in every transaction."

53. Blindness could have been improved. KCI should have taken additional precautions and established processes to minimize BellSouth's knowledge regarding impending and ongoing tests.

2. Military Style Testing.

54. Several of the concerns raised above are exacerbated by the manner in which KCI implemented military style testing. In a military style test, a

mindset of "test until you pass" is adopted. *See* Final Report at II-6. This approach can impact results.

55. First, military style testing greatly increases the chance of Type II errors. In other words, the military style third-party test is much more likely to conclude that BellSouth's process meets a test standard when it does not. Suppose that BellSouth provides chronically substandard service on a particular measure. As long as there is a possibility of Type II error on a single test, successive retests will eventually lead to a pass, resulting in a finding that BellSouth satisfied the standard when the result was merely the product of chance.

56. Second, KCI's military style test structure did not account for the increased scrutiny that should be applied in a retest. The fact that a measure failed one or more previous tests (a repeat offender) makes it more important to conduct a balanced analysis. This balanced analysis should include, among other things, increased sample sizes for retests. As KCI's witness acknowledged, "as sample size increases you should get closer to the right answer." Salzberg Tr. 91:3-5 (RMB-2). It is especially important to use larger for retests because the initial tests provided hard evidence of a problem that warrants close scrutiny. In a number of retests, however, KCI used sample sizes that were smaller than the original test where BellSouth failed. KCI's witness admitted that his understanding of the military-style test philosophy was that statistical tests are not redone each time: "it doesn't mean you have a wholesale retest." Salzberg Tr. 66:20-25 (RMB-2).

57. Examples where KCI retested smaller sample sizes include the following sixteen tests: PRE-1-3-1, PRE-1-3-2, PRE-1-3-3, PRE-1-3-4, PRE-1-3-5,

PRE-1-3-6, PRE-1-3-7, PRE-1-3-8, PRE-1-3-9, O&P-2-3-2a, O&P-2-3-2b, O&P-2-3-3b, O&P-5-1-1, O&P-5-2-1, O&P-5-2-2, and O&P-5-2-5.

58. For six of these measures—PRE-1-3-1, PRE-1-3-2, O&P-2-3-2a, O&P-2-3-2b, O&P-5-2-2, and O&P-5-2-5—KCI reached a conclusion of satisfied based only on the computed P-value. As noted earlier, the inadequate sample sizes for these six measures dictate against conclusions of satisfied. When also viewed in the context of BellSouth's initial test failures, these retest results clearly do not support the conclusion that BellSouth's true performance meets the standards.

59. KCI could have avoided this concern by retesting with significantly larger sample sizes.

c. Use of Professional Judgment to Overrule Observed Data.

60. I am concerned that so many measures that failed on the basis of the statistical analysis were reclassified as satisfied based on "professional judgment." For example, measure PRE-1-3-8, mean time for Service Availability Queries (SAQs), had a GPSC-approved standard of parity with retail performance, which was 1.3 seconds. The mean for pseudo-CLEC cases was 11.6 seconds, which was statistically significant in comparison to the BellSouth retail mean (*i.e.*, P-value < 0.05). However, KCI reversed this classification, concluding, "it is KCI's professional judgment that the average response interval for Test-CLEC-submitted SAQ pre-orders is within a reasonable timeframe." Similar reversals occurred for another twenty tests: PRE-1-3-3, PRE-4-3-1, PRE-4-3-2, PRE-4-3-3, PRE-4-3-4, PRE-4-3-5, PRE-4-3-8, PRE-5-3-1, PRE-5-3-2, PRE-5-3-3, PRE-5-3-4, PRE-5-3-5, PRE-5-3-8, O&P-5-1-1, O&P-10-3-5, O&P-10-3-6, O&P-10-3-7, O&P-10-3-8, O&P-10-3-9, and O&P-10-3-12.

61. Although I lack the business knowledge to comment on the validity of these judgments, I find it an unusual and questionable statistical practice to change evaluation criteria after seeing the data. If KCI thought that parity was too high a standard for certain measures, it should have specified revised criteria before doing the statistical analysis, not after it saw the results.

62. It is curious that KCI, after seeing the results, judged 11.6 seconds a reasonable average time for SAQs while, in its "professional judgment," setting 8.0 seconds as the standard for a long series of other measures. *See, e.g.*, PRE-1-3-6, PRE-1-3-7, PRE-1-3-9, PRE-4-3-6, PRE-4-3-7, PRE-4-3-9, PRE-5-3-6, PRE-5-3-7, PRE-5-3-9, O&P-10-3-10, O&P-10-3-11, and O&P-10-3-13.

63. KCI filed a Motion for Leave to Articulate Basis of Statistical Analysis in the Georgia 271 Test Final Reports in the Section 271 docket before the Louisiana Public Service Commission, Docket No. U-22252-E, attached as Exhibit RMB-3. KCI stated, "[a]s the author of the Georgia 271 Test Final Reports, no other party can adequately represent and articulate the basis for KPMG Consulting's use of statistical analysis in such reports." *Id.* ¶ 6.

64. None of KCI's comments refute my criticisms of the test design or KCI's conclusions. Indeed, KCI's discussion of the general design of the test acknowledges its highly subjective nature. KCI also admits that "the sample sizes for each specific service or transaction type were not designed for statistical precision."

65. In addition, KCI's filing fails to address a number of my points. For example, regarding the use of statistical analyses, KCI does not dispute my statement that the tests provide no information regarding how poor BellSouth's true performance is

likely to be in light of the observed data. While KCI's comments are technically correct on this issue, they do not address or refute my criticisms that the KCI statistical analysis is incomplete in this respect. Similarly, its defense of using statistical analysis for benchmark measures completely ignores the basis of my criticism. Both BellSouth's V-SEEM and the AT&T Performance Incentive Plan treat benchmarks as strict cutoffs for all samples of 30 or larger—implying that sufficient allowance has already been made for random variation.

66. Moreover, KCI's challenge to my concerns regarding inadequate retest sample sizes is simply illogical. The results from the smaller sample sizes used on retest cannot, as KCI claims, have been more focused and therefore more powerful than the original test. The data are what they are. They do not become more powerful by virtue of the purpose for which they were collected. As discussed above, samples of the sizes KCI used on retest are simply too small to rule out meaningful Type II errors; therefore they cannot form the basis for reliable conclusions.

IV. Performance Measures Evaluation

67. This section of my affidavit relates to BellSouth's performance measures plans in Georgia and Louisiana. Additional improvements are needed in order to evaluate whether BellSouth provides nondiscriminatory access to its network. First, AT&T believes that the remedy plan implements truncated z improperly by aggregating cells in a way that could conceal discrimination. Second, the parameter delta value should be lower to ensure parity. Third the remedy plan should calculate the parity gap in a way that penalizes BellSouth based on how far it stays from providing parity of performance.

A. Aggregation

68. Although truncated z is a valid method for aggregating cells, AT&T believes that the remedy plan implements truncated z improperly by aggregating cells in a way that could conceal discrimination.

69. Truncated z can allow parity service in some cells to conceal discrimination in other cells. The truncation step, setting $Z_j^* = \min(0, Z_j)$, is designed to keep a single cell where the CLEC's customers receive much better than parity service from canceling out poor service in other cells. However, it does not prevent parity, or better, service in a large number of cells from concealing very poor service in other cells. Suppose that in cells being aggregated BellSouth provides very poor service in a few cells (*e.g.*, modified z scores extreme enough to rule out random variation as the explanation) and parity service in other cells. The more parity cells that are included, the greater the chance is that truncated z will not be significant. The reason is that each cell that is found to be in parity increases the value of the truncated z statistic (high values are taken as evidence of parity). In addition, each new cell (whether in parity, or not) decreases the balancing critical value that truncated z must fall below to be judged significant. Similarly, parity service in just a few large cells can conceal very poor service in much smaller cells because truncated z weights the modified z scores according to sample sizes in the cells.

70. Consider a simple example with just two cells, using delta equal to 1.0. Assume that BellSouth provides a very large number of DS3 and POTS loops to itself with means and standard deviations of 5 days for each product. Now suppose that BellSouth provides a CLEC 30 DS3 loops in an average of 10 days and 250 POTS loops in an average of 5.1 days. The modified z for DS3 is –5.48, overwhelming evidence of

discrimination, and easily significant compared with the balancing critical value (BCV) of -2.74. The modified z for POTS is –0.32, which is not significant compared with a BCV of -7.90. If the two cells are aggregated using truncated z, the resulting truncated z score of –2.71 is much less extreme than the modified z for DS3 alone and is not close to significant when compared with the BCV of -10.24 for the aggregated test. Consequently, no remedy would be paid despite the clear evidence of large discrimination for DS3. Similar examples could easily be given for other values of delta.

B. Value of the Parameter Delta.

71. The parameter "delta" defines the degree of violation from parity for which the probability of Type II error is balanced against the probability of Type I error under parity. Delta specifies the difference between the CLEC mean and the BellSouth mean. In its General Order on Performance Measurements, the Louisiana Commission adopted the final staff recommendation of a delta of 1.00 for individual CLECs (Tier I) and 0.50 for CLEC Aggregate (Tier II) "for an interim period review period." (*See Staff Final Recommendation,* Attachment A to General Order in Louisiana Public Service Commission Docket U-22252, sub-docket C (May 14, 2001) at 12.) The Georgia Commission adopted values of 0.50 and 0.35 for Tier I and Tier II, respectively.

72. Delta should be set at the minimum value that represents a material impact on competition for a particular measure. AT&T believes that any value larger than 0.25 would not adequately protect CLECs against Type II errors. Accordingly, the Commission should adopt 0.25 or less as the parameter delta value for all submeasures in both Tier I and Tier II.

73. To understand the implications of delta = 0.25 and alternative values of delta, consider what they imply for an interval measure. Suppose that Order

Completion Interval for BellSouth customers has a mean of 5.0 days and a standard deviation of 5.0 days. Specifying delta sets the alternative hypothesis for which Type II error is balanced against Type I error. This alternative hypothesis states that the CLEC mean equals the BellSouth mean (5.0 days) plus a disparity of delta times the BellSouth standard deviation (delta x 5.0 days). Table 1 shows what this implies for three values of delta: 0.25, 0.50, and 1.00. A value of delta equal to 0.50 would be justified only if any disparity of less than 2.5 days is judged *not* to pose a material impact on competition. A delta of 1.00 would be justified only if any disparity of less than 5 days is judged *not* to pose a material impact on competition—*i.e.*, only if doubling the order completion interval was judged to be immaterial.

Table 1

Implied Disparity for Order Completion Interval,
by Value of Delta

	Delta		
Item	0.25	0.50	1.00
	(Days)		
Disparity [a]	1.25	2.50	5.00
CLEC mean under alternative hypothesis [b]	6.25	7.50	10.00

Table assumes the BellSouth mean and standard deviation are both 5.0 days.
[a] Disparity = delta x BellSouth standard deviation
[b] CLEC mean = BellSouth mean + disparity

74. Next, consider a counted measure indicating a particular service problem that is triggered for 1 percent of BellSouth's own customers. Column 1 of

Table 2 shows that the degree of disparity quantified by delta equal to 0.25 implies that 5.0% of CLEC customers would encounter the same problem; that is, the CLEC rate is five times the BellSouth rate.[4] Subsequent rows of the same column show the problem rates for CLEC customers implied by a delta of 0.25 for problems that affect 5, 10, or 20 percent of BellSouth customers. AT&T judges that disparities of this size pose material obstacles to competition. Therefore, delta should be no more than 0.25. Any larger value of delta would require even greater disparities before balancing takes place. For example, for a problem that occurs for 1 percent of BellSouth customers, a delta value of 0.50 would not balance the two types of error until the CLEC rate reached 11.8%, nearly a twelve-fold increase. For a delta value of 1.00, the two types of error wold not balance until the CLEC rate reached 31.9%, nearly a thirty-two-fold increase. These disparities are highlighted in Table 2.

[4] The table assumes use of the arcsine square root transformation to stabilize the variance of observed proportions. Using this function, transformed proportions have a nearly constant variance across the range of possible true proportions.

Table 2

Percentage of CLEC Customers Receiving Bad Service,
by BellSouth Percent and Delta

	Delta		
BellSouth Percent	0.25	0.50	1.00
1.0	5.0	11.8	31.9
5.0	11.8	21.0	44.0
10.0	18.7	29.3	53.6
20.0	30.8	42.8	67.4

75. Suppose that delta is set substantially above the minimum value that represents material impact on competition for a particular measure. Then the CLECs will face greater risk of a Type II error in the face of disparity constituting material impact than BellSouth would face of a Type I error under parity. In other words, proper balancing would not occur. This problem would be magnified for large sample sizes, because balancing can produce unconventionally large, negative critical values. For example, with samples sizes of 2,500 and 250 for BellSouth and a CLEC, respectively, a delta equal to 0.50 yields a balancing critical value of –3.77, corresponding to a Type I error probability of 0.00008 (*i.e.*, 1 in 12,000), far below any conventional significance level used in statistical testing. A delta equal to 1.00 would yield a balancing critical value of –7.54, corresponding to a microscopically small Type I error probability.

Consequently, compelling statistical evidence of discrimination, *e.g.*, a z score of -6.0, might be ignored. Such an outcome would be justified only if one could be certain that delta had not been set too large. Consequently, it is imperative not to set delta too large.

76. Dr. Mulrow states that materiality corresponds to a disparity of one-half delta times BellSouth's standard deviation. Mulrow Aff. ¶¶ 52-53. I base my definition on the principle behind balancing, that the probability of a Type I error assuming parity should equal the probability of a Type II error assuming a material disparity. Consequently, materiality refers to the size of the disparity specified in the alternative hypothesis—delta times the BellSouth standard deviation.

77. Including Dr. Mulrow's factor of one-half violates the basic principle of the balancing critical value methodology. Balancing occurs when the true difference in means equals delta x BellSouth's standard deviation. The Louisiana joint statistician's report implicitly defines materiality in terms of the alternative hypothesis, "If a standard of materiality is set by stating a specific alternative hypothesis for the test, …then a critical value can be determined so that the two error probabilities are equal." (Mulrow Aff., Exhibit EJM-1, p. 8). That is, a material difference must be defined as delta x BellSouth's standard deviation (the difference between the BellSouth mean and the CLEC mean under the alternative hypothesis). If delta is set incorrectly, so that a difference of one-half that size is material, then proper balancing does not occur. The probability of a Type II error when there is a difference corresponding to one-half delta remains at 50 percent, no matter how low the Type I error falls.

c. Affected Volume.

78. The calculations illustrating the SEEM remedy procedure (pp. 38-41 of Varner's Exhibit PM-8) are incorrect. Although the ILEC sample sizes for cells 1-

10, which are not provided, would be required to validate the modified z and truncated z values, there is enough information available to prove that the balancing critical values shown in the tables are wrong by as much as a factor of 70. The tables all report balancing critical values of –0.21. However, for Order Completion Interval (*id.* 40), if the total ILEC sample size of 50,000 is divided equally among the ten cells, the correct balancing critical value (BCV) is -14.58. If, instead, the ILEC sample is divided in proportion to the CLEC sample, the correct BCV is -14.67. Even if each ILEC cell size were only 10 (for a total of ILEC sample of 100), the correct BCV would be -4.75. Under any of these three scenarios for the correct BCV, a truncated z of –1.92 would not even approach the BCV, and no payout would be made. Consequently, all three tables give a distorted impression of the SEEM remedy procedure.

79. The remedy calculations also should be improved. As currently set, BellSouth may stray far from providing parity with only limited consequences. Absent meaningful consequences, BellSouth has little incentive to provide parity.

80. The remedy calculation the Commission has adopted for retail-analog measures multiplies "per affected item" dollar amounts by a calculated "affected volume." There is no justification that this calculated affected volume produces any semblance of a true affected volume. Indeed, the so-called "parity gap," which is a direct factor in the formula for affected volume, clearly calculates something other than its name implies. Instead of computing how far BellSouth was from providing parity service, the parity gap computes how far BellSouth was from not being found in violation.

81. Consider this analogy. Suppose that the police patrol a stretch of highway with a 65-MPH speed limit, but that they only stop drivers who exceed 75 MPH. Also, suppose that state law calls for a fine of $10 per MPH in excess of the limit. If I am caught going 77 MPH, can I expect only a $20 fine because I was going just 2 MPH too fast to get caught? Unlikely. But that is analogous to how the plan computes remedies.

82. Although the statistical tests need to allow some leeway for random variation, we should not forget the goal of the Telecommunications Act is parity service. The current plans in Louisiana and Georgia allow BellSouth to stray far from the goal of parity, but suffer only minimal financial consequences. Accordingly, BellSouth's remedy plans will not be sufficient to prevent BellSouth from backsliding.